Monocular Underwater Vision Pipeline for 6DoF Annotations with Inpainting-Based Image Augmentation

Alexander Klein[a], David Brandt[a], and Jannis Stoppe[a]

[a]German Aerospace Center (DLR), Institute for the Protection of Maritime Infrastructures, Bremerhaven, Germany

ABSTRACT

The acquisition of high-fidelity, annotated data for training perception and manipulation tasks poses significant challenges. This process typically demands customized setups, tightly controlled environments, and specialized sensing equipment that are unavailable in underwater settings. Marker-based methods offer a simpler alternative by tracking the six degrees of freedom poses of objects using a monocular camera. However, attaching markers to objects alters their original form and appearance, while placing markers in the environment modifies the backdrop and limits the flexibility and portability of such methods.

In this work, we present a pipeline capturing underwater scenes using a pose plate with fixated featureless objects of varying scales. The pose plate is equipped with ArUco markers, which track the 6D camera pose and enable the pipeline to render pixel-wise depth and object masks. Custom camera mappings ensure precise alignment between rendered masks and sensor images. To prevent machine learning models from relying on the markers as cues rather than building robust object representations, our pipeline employs object aware inpainting as augmentation method, replacing the pose plate with a realistic background.

The pipeline was validated by training semantic segmentation models on a custom dataset consisting of scenes in different underwater environments. Our experiments demonstrate that incorporating augmented data into the training process yields improved model performance, outperforming models trained solely on images with visible markers. This finding suggests that our proposed techniques have the potential to mitigate the domain gap between marker-based ground truth and real-world data.

Keywords: Underwater Perception, 6D Object Pose, Image Augmentation, Semantic Segmentation, Computer Vision, Underwater Dataset

1. INTRODUCTION

Due to the inherent risks associated with underwater operations, remotely operated vehicles (ROVs) have become an increasingly attractive solution for improving diver safety and potentially replacing human divers in certain applications. However, a key challenge in operating ROVs is their limited dexterity, particularly in complex manipulation tasks, such as grasping. While research on autonomous manipulation tasks is well-established for industrial and mobile robots, with numerous ongoing efforts for six degrees of freedom (6DoF) pose estimation[1,2] and grasp synthesis,[3,4] the literature on this topic in the underwater environment is relatively sparse.[5]

Existing datasets for 6DoF pose estimation often overlook the underwater environment, as they typically require specialized settings or sensors that are not tailored for underwater applications, necessitating either custom-designed sensors or controlled environments to collect data. Datasets such as LineMod[6,7] rely on RGBD cameras and utilize the geometry of known objects, while others, like the HOPE[8] dataset, employ hand-annotated key-points for pose annotation and HOT3D[9] relies on a complex multi-camera setup. These datasets are primarily recorded indoors, focusing on grasping within the workspace of 6DoF robot arms. In contrast, outdoor scenes for autonomous driving are acquired using LiDAR sensors to estimate depth in the KITTI[10] dataset, with data typically collected from the vantage point of a moving vehicle.

Further author information: (Send correspondence to A. Klein)
A. Klein: E-mail: alexander@dlr.de

Applications of Machine Learning 2025, edited by Michael E. Zelinski, Tarek M. Taha, Barath Narayanan Abdul A. S. Awwal, Khan M. Iftekharuddin, Proc. of SPIE Vol. 13606, 136060Q
2025 Published by SPIE · 0277-786X · doi: 10.1117/12.3063565

The majority of these methods are based on precise depth data, often employing LiDAR systems or depth-sensing cameras that utilize infrared radiation to obtain accurate measurements. However, as electromagnetic waves of the infrared spectrum travel through water, they experience significant energy loss due to absorption, scattering, and the presence of particles, resulting in a weaker return signal with increasing distance.[11,12] Despite efforts to adapt conventional depth cameras for underwater 3D reconstruction, these attempts have been hindered by significant sensor noise, rendering them reliable only at short ranges of up to 0.8 meters, even under ideal water conditions.[13,14]

To address these challenges, we propose a vision pipeline for collecting and annotating 6DoF pose data with semantic object information using a monocular camera. Building upon existing methodologies[15] we employ fiducial markers for pose estimation, but extend the approach using a plate with multiple markers as basis for our scene. To mitigate the introduction of artificial elements into the scene, we utilize object-aware inpainting as a data augmentation technique to seamlessly remove the fiducial markers from the image, ultimately generating a more realistic and challenging environment that resembles real-world conditions. We leveraged our pipeline to collect a small underwater dataset, which we plan to make publicly accessible. By training machine learning models on this dataset for semantic segmentation, we aimed to validate the efficacy of our augmentation method in bridging the gap between marker-based acquisition and real-world data.

2. METHOD

2.1 Data Acquisition Pipeline

The goal of the pipeline is to acquire instance level semantic, depth and 6D pose annotations given a monocular camera and predefined target objects. For this we assume that all objects used in the data set are known beforehand and that their geometry is available as a 3D mesh. Furthermore we require the intrinsic camera parameters as well as the parameterized distortion model calibrated for the right medium, as the refractive index in water $n \approx 1.33$ results in a proportional longer focal length.[16] Besides the intrinsic ones, this affects other relevant lens parameters as well. As wide angle and ultra wide angle (fish eye) lenses are popular tools for under water operations, the distortion model has to be calibrated for optimal results. The 6D pose estimation is based on the idea that each target object is known beforehand and placed on a pose plate which will be tracked using fiducial markers. The pose plate defines the base coordinate system and its orientation, every other object must either resolve the transformation from itself to the pose plate or has to be hard mounted at fixed position and orientation. After altering the camera viewport or scene (e. g. by manually rearranging objects) the 6D pose of all objects and sensors must be updated by resolving the kinematic chain in the scene graph. We track the 6D pose of the camera in relation to the pose plate by detecting corner points of multiple ArUco markers.[17,18] As the exact relative locations of these points is known beforehand we can resolve the 6D Transformation from world (pose plate) to camera frame given all detected corners in image coordinates using the Perspective-n-Point (PnP) algorithm.[19] In comparison to solving the PnP with 4 corners for a single ArUco marker this approach allows for more stable results, using RANSAC[20] to avoid potential outliers.

Using the pose graph with known object meshes and the tracked camera we create a virtual scene for rendering, where the exact camera parameters and distortion parameters are used to generate accurate virtual to reality pixel to pixel mappings. For the rendering process we perform a separate render pass for each object, generating exact depth and annotation masks. This separation allows multi modal training objectives as they can be reconstructed for task specific training as explained in section 2.3.

2.2 Data Pruning

A limitation of the method described in Section 2.1 is that the accuracy of the estimated 6D pose is dependent on the accuracy of fiducial marker pose evaluation, which can be decomposed into translation and rotation errors. As the accuracy of the translation might vary, it mostly hinges on the estimation of the depth estimate (z-axis in OpenCV camera coordinates) as the other dimension are inferred directly by the projection of the marker on the camera sensor. The rendered object masks align closely with the corresponding recorded images, as their projections coincide with the corresponding object silhouettes. Although inaccurate depth estimates may result in slightly smaller or larger projections, this effect is typically negligible due to the projection's dependence on

the absolute distance between object and camera, which is often several meters. Consequently, even errors of multiple centimeters have a relatively minor impact on the pixel-wise mapping from rendered to raw image.

More substantial challenges emerge from uncertainties associated with estimating rotation parameters, which become increasingly pronounced as visibility deteriorates due to factors such as haze, turbidity, or low light conditions underwater.[21] Even slight inaccuracies can have a profound effect on the dataset, causing distortions in the object's silhouette and resulting in misalignment in the pixel-wise mapping. As we rely on the markers to build a ground truth we are lacking any mechanism to automatically correct this error.

To enhance the accuracy of our automatic method, we introduced a visual inspection step, which compromises the pipeline's full automation. A human operator reviews the alignment of rendered silhouettes with the target objects and manually prunes outliers from the dataset. This step enables a tighter alignment between the virtual and recorded data, leveraging human visual inspection to verify accuracy without requiring manual annotation of specific values, minimizing the workload for annotations by hand. A straightforward approach to mitigate over-pruning is to collect additional data, as recording more data is relatively inexpensive compared to manual annotation by humans.

2.3 Post Processing

To generate other modalities of data, the pipeline generates pixel-wise mapping from sensor to virtual space for depth maps as well as the option for rendered annotations embedded as texture for each mesh individually. The primary task addressed in this paper – semantic segmentation – relies solely on object masks derived from depth data. As each mesh is rendered separately we omit the evaluation of intersections in the depth buffer in the rasterization pipeline. Task-specific annotations that necessitate customized occlusion handling or rely on selected objects are typically either precomputed during a post-processing step or managed directly by a specialized data loader. This approach enables occlusion handling to be seen as additional parameter and opens the choice to select a subset of target objects after the dataset creation and make use of occluded object masks.

Given N pixel aligned depth maps $\mathbf{I}_n$ with height i and width of j, we can estimate the complete depth map $\mathbf{D}$ by taking the element-wise minimum, as described in Equation (1). Additionally we define the validity mask $\mathbf{M}$ in Equation (2) which indicates the presence of a valid depth value (denoted by 1) at each pixel location.

$$\mathbf{D}_{xy} = \min_{n=1,\dots,N}(\mathbf{I}_{nxy}) \quad \text{for } x = 1,\dots,i \text{ and } y = 1,\dots,j \tag{1}$$

$$\mathbf{M}_{xy} = \begin{cases} 1 & \text{if } D_{xy} > 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } x = 1,\dots,i \text{ and } y = 1,\dots,j. \tag{2}$$

In analogy to the depth map $\mathbf{D}$ we can create a segmentation map $\mathbf{S}$ evaluating the class ID n of the minimum depth value at spatial pixel location i, j given the aligned depth maps $\mathbf{I}_n$ as seen in Equation (3).

$$\mathbf{S}_{xy} = \arg\min_{n=1,\dots,N}(\mathbf{I}_{nxy}) \quad \text{for } x = 1,\dots,i \text{ and } y = 1,\dots,j \tag{3}$$

2.4 Inpainting

A potential issue using fiducial markers to generate training data stems from the fact that they are highly visible image features that offer at least a partial deterministic solution solving the 6DoF pose estimation. Furthermore, depending on the training setup, a machine learning model may exploit these external features such as visible markers and the pose plate as visual cues to infer information about target objects, rather than relying solely on the relevant object characteristics.

In real world applications the environment is unordered and doesn't provide any visual guidance and may have one or multiple target objects in unseen orientations. This domain gap between annotated data and reality might pose problems transferring results into the real world. Since the ground truth data inherently relies on

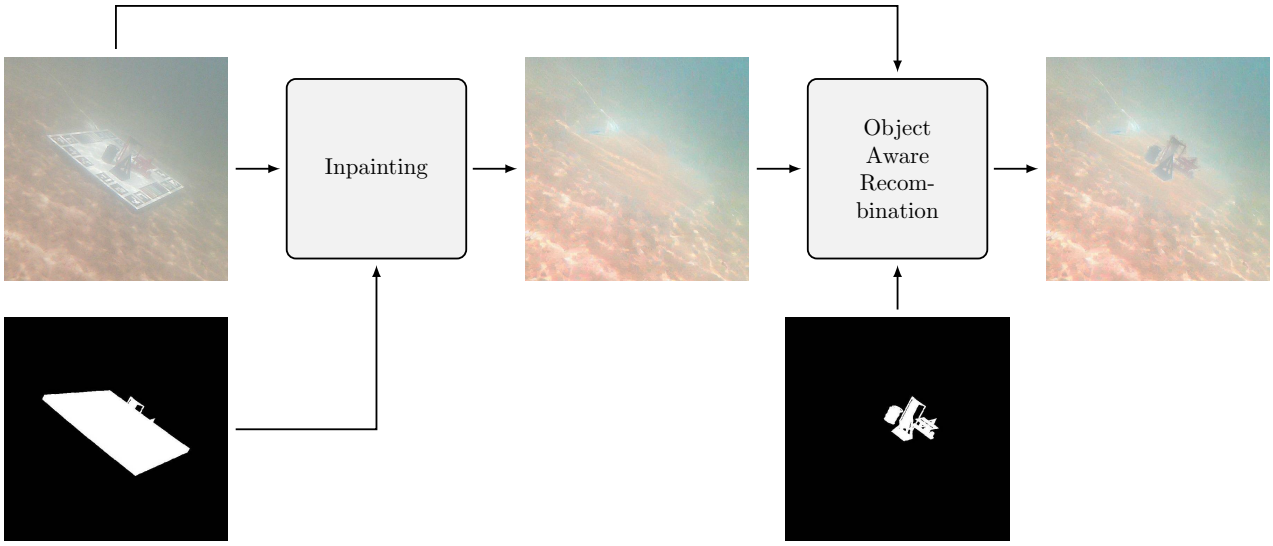
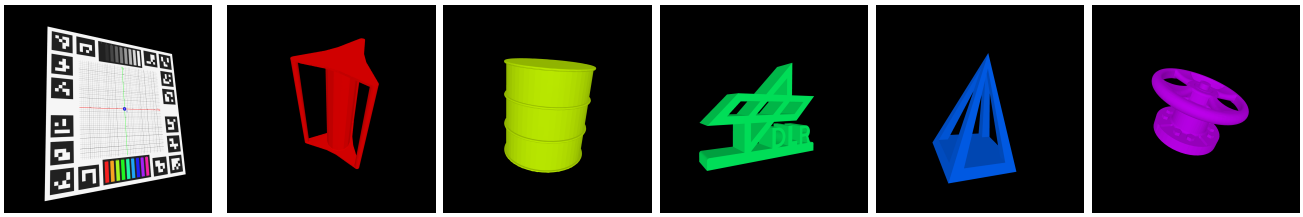


Figure 1: Complete inpainting pipeline for generating augmented data. The process involves two steps: (1) stable diffusion based inpainting of all objects with realistic background, and (2) alpha blending to reintroduce the target objects into the frame, effectively removing the pose plate with fiducial markers.



(a) Pose Plate (b) Grasp Body (c) Barrel (d) Logo (e) Pyramid (f) Valve

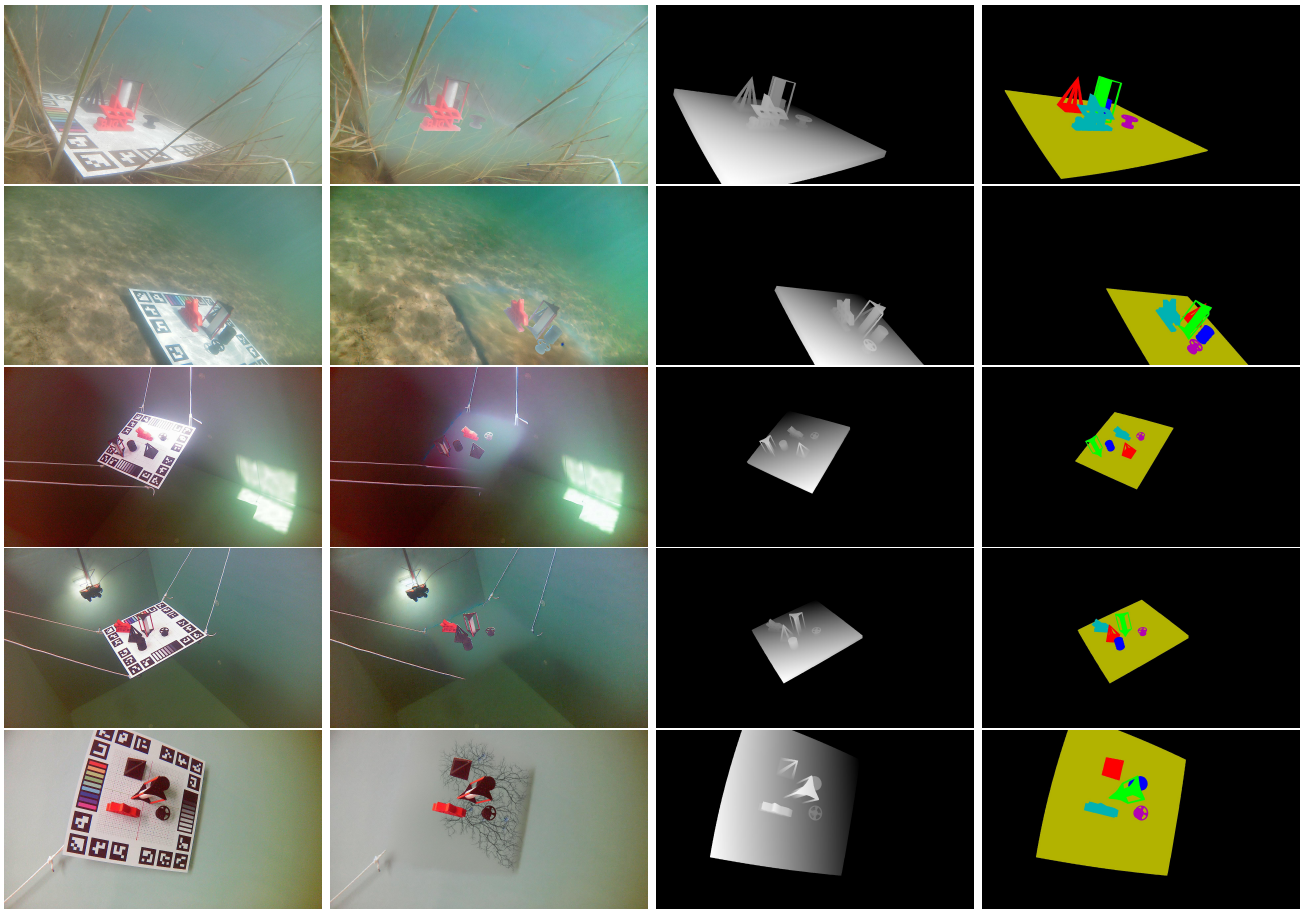
Figure 2: Meshes of pose plate and target objects used in the pipeline to create the virtual scene.

the presence of markers, the characteristic is shared across the training, validation, and test sets. This raises concerns when comparing the performance of trained models, as a model that performs better on the test set may still underperform in real-world scenarios if its success largely results from exploiting marker information rather than learning generalisable features.

Due to this issue we augment the original image by removing the pose plate with the fiducial marker from the image without editing the target objects. We employ mask aware inpainting using stable diffusion[22] to fill in object masks with background information as augmentation method. The complete process with intermediate results is shown in Figure 1. To avoid hallucinations in the diffusion process, which can occur when the geometric structures of the target objects are complex, we adopt a 2-step approach instead of simply removing the pose plate mask in a single step. For the first step, the inpainting process, we generate one hot encoded object mask using all objects in the scene as described in Equation (2). Using the original sensor image, stable diffusion generates a background image by filling in the masked area, effectively removing the pose plate and all target objects. In the next object aware recombination step, we utilize a new mask that excludes the pose plate and use alpha blending[23] to transfer the appearance of the target objects from the sensor image to the augmented background image. Compared to purely synthetic data this augmentation method preserves environmental influences on the objects, e.g. lighting, turbidity, caustics and hazing.

3. DATASET

Our dataset consists of 9298 recorded images based on two underwater environments. Around 80 % were taken in a deep saltwater pool with depth of roughly 5 meters and the other 20 % were collected in a freshwater lake. The target objects were selected based on their surface complexity and potential as training targets for future



(a) Sensor Image (b) Augmented Image (c) Depth Map (d) Semantic Segmentation

Figure 3: Examples from our dataset, illustrating the available data modalities. Shown are: (a) raw images, (b) augmented images via inpainting, (c) depth maps, and (d) semantic segmentation maps. The top two rows display samples from a freshwater lake environment, while the remaining images were taken in a deep saltwater pool.

underwater grasping tasks. The sensor employed is a single, waterproof, monocular camera equipped with a fisheye lens, providing an approximate horizontal underwater field of view of 82° with a sensor resolution of 1920×1080 pixels.

All 5 3D printed target objects as well as the pose plate with the fiducial makers are shown in Figure 2. The grasp body, barrel, pyramid and valve are all bodies of revolution, with the barrel being the simplest structure and the others exhibiting fixed symmetries along the rotational axis. In contrast, the logo is a more complex object, yet still possesses a symmetric axis.

For the data acquisition various scenes were set up under varying lighting conditions. Images were primarily captured from orbital positions in the upper hemisphere of the pose plate, at distances ranging from 0.5 to 5.5 metres. Example images from the lake and pool environments, along with augmented images and visualizations of depth and semantic segmentation maps generated through post-processing (described in Section 2.3), are shown in Figure 3. Following the manual pruning described in subsection 2.2, 3910 high quality annotated images are left with available image augmentation, segmentation and depth maps. The dataset is further subdivided into 3,128 training samples, 391 validation samples, and 391 test samples.

4. VERIFICATION AND SEMANTIC SEGMENTATION RESULTS

To mitigate the risk of a model relying on the pose plate and the attached fiducial markers as a cue for machine learning tasks, potentially limiting its generalizability to real-world scenarios, we developed an inpainting based data augmentation method for the data generation pipeline. Notably, since the pose plate itself provides the necessary ground truth for pose estimation, our dataset does not contain any non-augmented images without the pose plate. Therefore we are not able to directly verify that our augmentation method changes the image space enough to close the domain gap to realistic scenarios. We instead employ a simple cross-validation scheme for a basic semantic segmentation task, where the goal is to segment the objects *grasp body*, *barrel*, *logo*, *pyramid* and *valve* to analyze the impact of the augmentation method. All other pixels, including those belonging to the pose plate, are assigned to the *other/background* class. We differentiate between three input scenarios: (1) **raw** sensor data, (2) **aug**mented images, and (3) the **full** dataset, each providing identical segmentation masks as target for the training. To investigate the impact of each scenario on model performance, we trained simple segmentation models on each corresponding training and validation set. For a comprehensive cross-comparison, we utilized separate test sets for each dataset and evaluated the performance of every model on each respective dataset.

4.1 Models

For the machine learning model we use the decoder of the dense prediction transformer (DPT)[24] architecture which allows the stage wise recombination of features generated by an arbitrary feature pyramid backbone. Since our primary objective was to analyze cross-correlations rather than achieve optimal performance on the dataset, we employed simple convolutional backbones, specifically ResNet-34 and ResNet-50.[25] Notably, combining ResNet-50 with the DPT decoder corresponds to the DPT-Hybrid[24] architecture. While the decoders' weights are initialized using random values the ResNet backbones use pretrained parameters[26,27] based on the ImageNet[28] dataset. We configured the model to accept images and produce segmentation maps of size 512 $\times$ 512 pixels. The ResNet-50 backbone utilizes a decoder with a feature depth of 128, whereas the ResNet-34 backbone employs a smaller decoder with a feature depth of 96, resulting in an overall smaller architecture.

4.2 Training Procedure

To ensure comparable results, we maintained a consistent training procedure across all models and datasets. We trained each model for 300 epochs using the AdamW[29] optimizer with a constant learning rate of $1e^{-5}$, weight decay of 0.01 and a batch size of 16. During the training process, we apply data augmentation techniques to enhance the diversity of our dataset. To achieve the input resolution we randomly utilize either resizing or random cropping with a simple normalization step as input augmentation method. Furthermore, we also incorporate random rotations to introduce variability in the orientations of the images.

The small size of the target objects results in a substantial class imbalance between background and annotated pixels. To address this issue, we employ a balanced binary cross-entropy loss function that adaptively scales the gradients based on the pixel distribution between annotated pixels and the background. By weighting the gradients inversely proportional to the number of pixels per class, we reduce the impact of the dominant background class. To avoid numerical instabilities caused by extremely large gradients due to small objects, we cap the scaling factor at 1000.

4.3 Results

Figure 4 shows the results of the ResNet-50 based architecture trained on the full dataset, which includes both raw and augmented images. The model demonstrates successful segmentation of even small objects at large distances, while maintaining their rough outline. Analyzing the results we observed that the model has a tendency that a segmented object silhouette produces a fuzzy outline labeling background as object pixels in neighbouring regions, often filling small details such as the holes in the silhouette of the valve, logo and pyramid. Since this effect predominantly occurs in the direction from target object to background, with barely any misclassification of object pixels as background, we suspect that small misalignments in the ground truth data prevent the model from learning a detailed and accurate object shape for its internal representation.

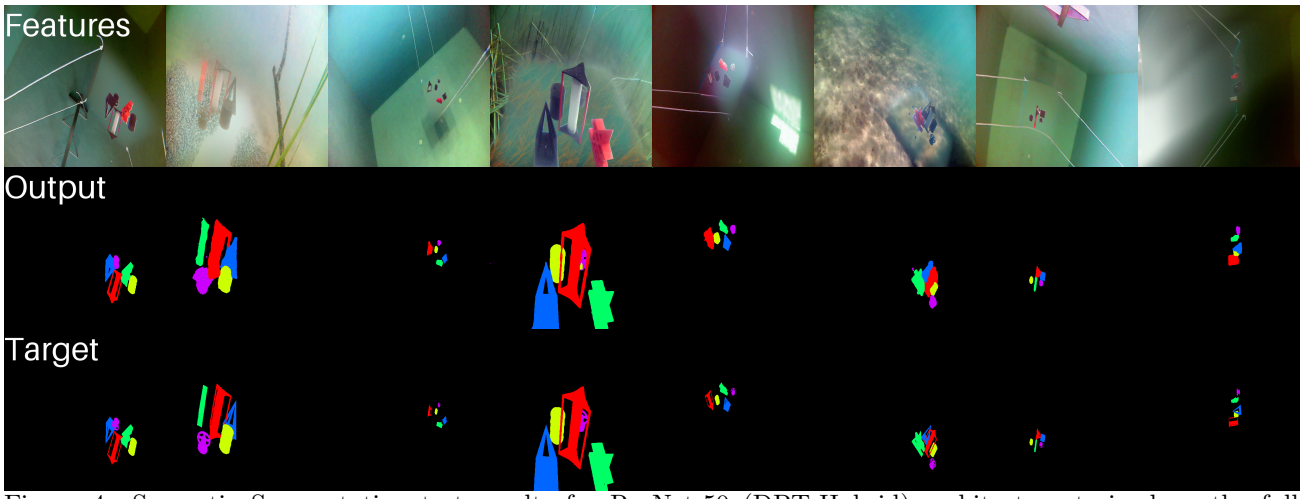


Figure 4: Semantic Segmentation test results for ResNet-50 (DPT-Hybrid) architecture trained on the full dataset.

Table 1: Results of the cross evaluation on the test set for Raw, Augmented (Aug) and Full dataset input modalities for the ResNet-34 and ResNet-50 based DPT-Hybrid architectures.

		mIoU			Precision		
Backbone	Trainingdata	Raw	Aug	Full	Raw	Aug	Full
ResNet-34	Raw	0.6451	0.5866	0.6156	0.6596	0.6365	0.6476
ResNet-34	Augmented	0.5461	0.7516	0.6354	0.6186	0.7715	0.6890
ResNet-34	Full	0.6714	0.7376	0.70324	0.6849	0.7544	0.7173
ResNet-50	Raw	0.6846	0.6051	0.6365	0.6960	0.6593	0.6767
ResNet-50	Augmented	0.4520	0.7849	0.5463	0.4700	0.7974	0.5624
ResNet-50	Full	0.6908	0.7727	0.7287	0.7006	0.7843	0.7393

Table 1 presents the mean Intersection over Union (mIoU) and Precision for cross-evaluation of dataset feature modalities, including raw (sensor images), aug (augmented images), and full (sensor + augmented images). Both ResNet-34 and ResNet-50 architectures trained on raw and augmented data perform worse on the other dataset, despite unchanged target and object appearance. Notably, ResNet-50 trained on augmented data experiences a 0.33 mIoU drop when tested on raw data, whereas training on raw data and testing on augmented data results in only a 0.08 mIoU loss. We observed that models trained on augmented data struggle to differentiate between markers and featureless geometric silhouettes, resulting in false positives along the markers on the pose plate borders. In contrast, models trained on raw data and evaluated on augmented data often incorrectly estimate segmentation masks, missing smaller objects like the valve, or in severe cases, mislabel all objects as background. The models trained on the full dataset demonstrate improved performance on raw data compared to dedicated models, while also almost matching the best results on augmented data category. This suggests that applying our inpainting augmentation method might improve model performance even for a general case without requiring additional training data. Our findings suggest that visible markers can compromise model performance, whereas our object-aware inpainting augmentation method can help bridge the domain gap, thereby improving robustness in real-world applications.

5. CONCLUSION

In this work, we introduced a versatile pipeline that leverages fiducial markers to enable semi-automatic generation of 6D pose, semantic, and depth maps using a single monocular camera, with potential applications extending beyond these tasks. Notably, our pipeline showcased the effectiveness of mask-aware inpainting as a data augmentation technique, utilizing precise object masks to enhance model robustness. Furthermore, we created a novel underwater dataset using our pipeline, which provides a unique and diverse range of data modalities for various tasks in an underrepresented environment. In our experiments we evaluated machine learning models,

comparing augmented and raw data images with the conclusion that the object aware inpainting offers a significant alteration to the feature space, which could potentially bridge the gap to realistic application scenario. Additionally, our experiments have demonstrated the success of object-aware inpainting as an augmentation method, which improves model performance.

Future work will focus on expanding the diversity of objects, locations and sensors within our dataset, enabling more comprehensive and robust evaluations. A lab setup with exact measurements not based on fiducial markers would enable a more detailed comparison of the proposed augmentation method and markerless approaches. To acquire non-synthetic depth data, we intend to leverage the pipeline's open formulation of the scene graph, which enables seamless integration of additional sensors. Notably, our plans include integrating a multibeam sonar, a underwater acoustic sensor, to facilitate sensor fusion research in the underrepresented domain of underwater perception.

ACKNOWLEDGMENTS

This research was conducted as part of the MUM2 project, which is funded by the German Federal Ministry for Economic Affairs and Energy (BMWE). The authors would like to acknowledge the financial support provided by the BMWE. Additionally, we would like to thank the Alfred-Wegener-Institut for providing access to their sea water pool and for their assistance in recording the dataset.

REFERENCES

[1] Tekin, B., Sinha, S. N., and Fua, P., "Real-time seamless single shot 6d object pose prediction," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 292–301 (2018).

[2] Xu, Y., Lin, K.-Y., Zhang, G., Wang, X., and Li, H., "Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 14880–14890 (2022).

[3] Bohg, J., Morales, A., Asfour, T., and Kragic, D., "Data-driven grasp synthesisa survey," *IEEE Transactions on robotics* **30**(2), 289–309 (2013).

[4] Newbury, R., Gu, M., Chumbley, L., Mousavian, A., Eppner, C., Leitner, J., Bohg, J., Morales, A., Asfour, T., Kragic, D., et al., "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics* **39**(5), 3994–4015 (2023).

[5] Huang, H., Tang, Q., Li, J., Zhang, W., Bao, X., Zhu, H., and Wang, G., "A review on underwater autonomous environmental perception and target grasp, the challenge of robotic organism capture," *Ocean Engineering* **195**, 106644 (2020).

[6] Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., and Lepetit, V., "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in [*2011 International Conference on Computer Vision (ICCV 2011)*], 858–865, IEEE, Piscataway, NJ (2011).

[7] Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., and Rother, C., "Learning 6d object pose estimation using 3d object coordinates," in [*Computer vision - ECCV 2014*], Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., eds., *Lecture Notes in Computer Science* **8690**, 536–551, Springer, Cham (2014).

[8] Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., and Birchfield, S., "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," (2022).

[9] Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J. J., and Hodan, T., "Introducing hot3d: An egocentric dataset for 3d hand and object tracking," (2024).

[10] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R., "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research* **32**(11), 1231–1237 (2013).

[11] Morel, A., Gentili, B., Claustre, H., Babin, M., Bricaud, A., Ras, J., and Tieche, F., "Optical properties of the clearest natural waters," *Limnology and oceanography* **52**(1), 217–229 (2007).

[12] Asano, Y., Zheng, Y., Nishino, K., and Sato, I., "Depth sensing by near-infrared light absorption in water," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(8), 2611–2622 (2021).

[13] Anwer, A., Ali, S. S. A., Khan, A., and Mériaudeau, F., "Underwater 3-d scene reconstruction using kinect v2 based on physical models for refraction and time of flight correction," *IEEE Access* **5**, 15960–15970 (2017).

[14] Dancu, A., Fourgeaud, M., Franjcic, Z., and Avetisyan, R., "Underwater reconstruction using depth sensors," in [*SIGGRAPH Asia 2014 Technical Briefs*], *SA '14*, Association for Computing Machinery, New York, NY, USA (2014).

[15] Sapienza, D., Govi, E., Aldhaheri, S., Bertogna, M., Roura, E., Pairet, ., Verucchi, M., and Ardn, P., "Model-based underwater 6d pose estimation from rgb," *IEEE Robotics and Automation Letters* **8**(11), 7535–7542 (2023).

[16] Lavest, J.-M., Rives, G., and Lapresté, J.-T., "Underwater camera calibration," in [*Computer VisionECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*], 654–668, Springer (2000).

[17] Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., and Marín-Jiménez, M. J., "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition* **47**(6), 2280–2292 (2014).

[18] Romero-Ramirez, F. J., Muñoz-Salinas, R., and Medina-Carnicer, R., "Speeded up detection of squared fiducial markers," *Image and vision Computing* **76**, 38–47 (2018).

[19] Fischler, M. A. and Bolles, R. C., "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in [*Readings in Computer Vision*], Fischler, M. A. and Firschein, O., eds., 726–740, Morgan Kaufmann, San Francisco (CA) (1987).

[20] Fischler, M. A. and Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* **24**, 381395 (June 1981).

[21] Risholm, P., Ivarsen, P. Ø., Haugholt, K. H., and Mohammed, A., "Underwater marker-based pose-estimation with associated uncertainty," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 3713–3721 (2021).

[22] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., "High-resolution image synthesis with latent diffusion models," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 10684–10695 (June 2022).

[23] Porter, T. and Duff, T., "Compositing digital images," *SIGGRAPH Comput. Graph.* **18**, 253259 (Jan. 1984).

[24] Ranftl, R., Bochkovskiy, A., and Koltun, V., "Vision transformers for dense prediction," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 12179–12188 (2021).

[25] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[26] Wightman, R., "PyTorch Image Models," (2019).

[27] Wightman, R., Touvron, H., and Jégou, H., "Resnet strikes back: An improved training procedure in timm," *CoRR* **abs/2110.00476** (2021).

[28] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (2009).

[29] Loshchilov, I. and Hutter, F., "Fixing weight decay regularization in adam," *CoRR* **abs/1711.05101** (2017).